University of Bedfordshire

Cognitive validity in language testing: theory and practice

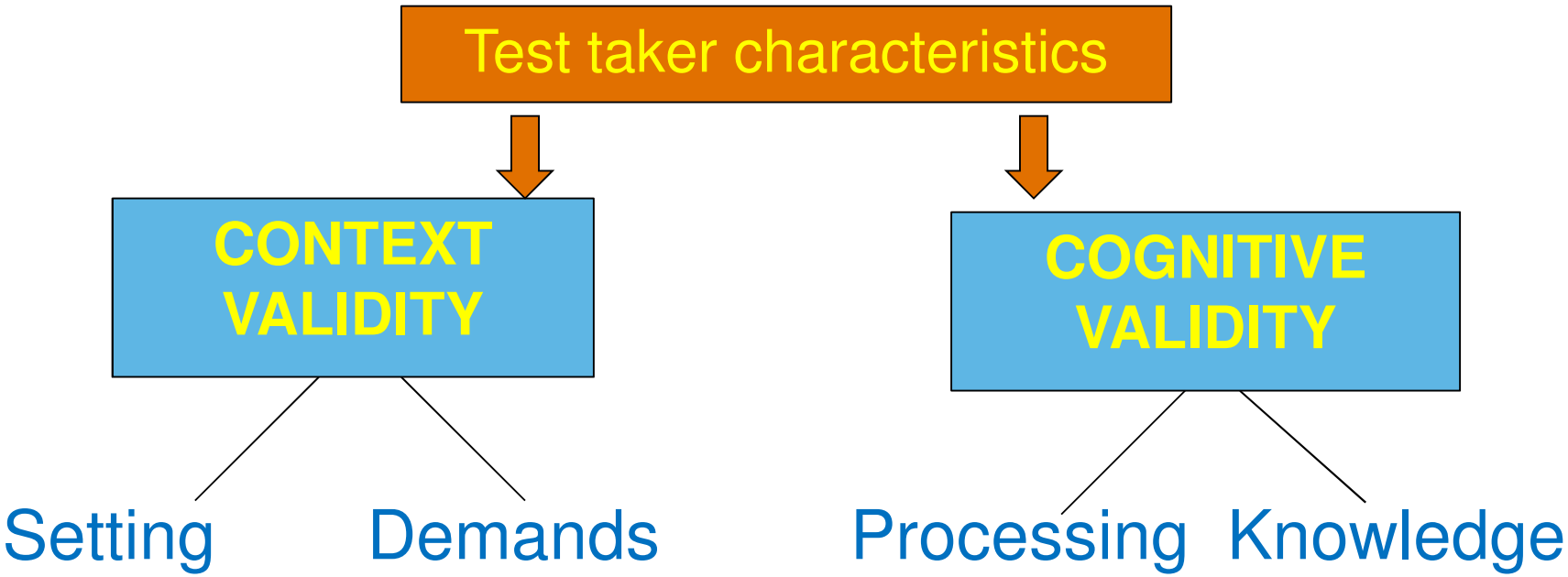
Dr John Field, CRELLA

University of Bedfordshire

Background

The socio-cognitive framework (Weir, 2005)

University of Bedfordshire



Cognitive validity (Glaser, 1991)

University of Bedfordshire

- The issue is not 'authenticity'. Clearly we cannot reproduce the circumstances of a real language event in the artificial environment of a test.
- But cognitive validity requires us to find out if the mental processes that a test elicits from a candidate resemble the processes that he/she would employ in real-world conditions.
- At issue: *How valid is the test as a predictor of real-life performance?*
- The notion of cognitive validity has been used to investigate whether tests of scientific thinking or logical reasoning actually tap in to the processes they are supposed to measure (rather than, e.g. relying on rote learned facts). Baxter & Glaser, 1998, Thelk & Hoole, 2006

Predictive testing

University of Bedfordshire

- Many high-stakes language test scores are employed predictively: e.g. to show that an individual is capable of performing in a particular job, class or academic setting.
- This places a responsibility on the test designer to ensure that the test elicits behaviour similar to the behaviour that happens in a real-world context.

Expertise

University of Bedfordshire

- An expert employs a skill in a way that is rapid and that does not demand forethought.
- A good driver does not have to think about the process of changing gears.
- A good L2 speaker constructs and produces a sentence without having to pause to think about the words or grammar being used.
- Expertise concerns how the knowledge stored in a performer's mind is a) accessed and b) put to use (i.e. it concerns procedural knowledge)

Cognitive validity and test design

University of Bedfordshire

- Weir (2005) argues that we need a clearer idea of the construct we aim to test before designing a test.
- Post-hoc validation of test results may tell us how well a test discriminates between candidates. Factor Analysis may tell us what traits the test taps into.
- But we cannot link this to construct validity unless we fully understand the construct we are trying to measure. e.g. Which traits contribute importantly to the skill? Which do not?

Establishing cognitive validity

University of Bedfordshire

- Cognitive validity can be investigated in two ways:

1. How does an expert language user behave (what is the target behaviour learners are working towards?)

→ Modelling the skill

2. What do test takers actually do in a test? How closely does it resemble behaviour in real-world contexts?

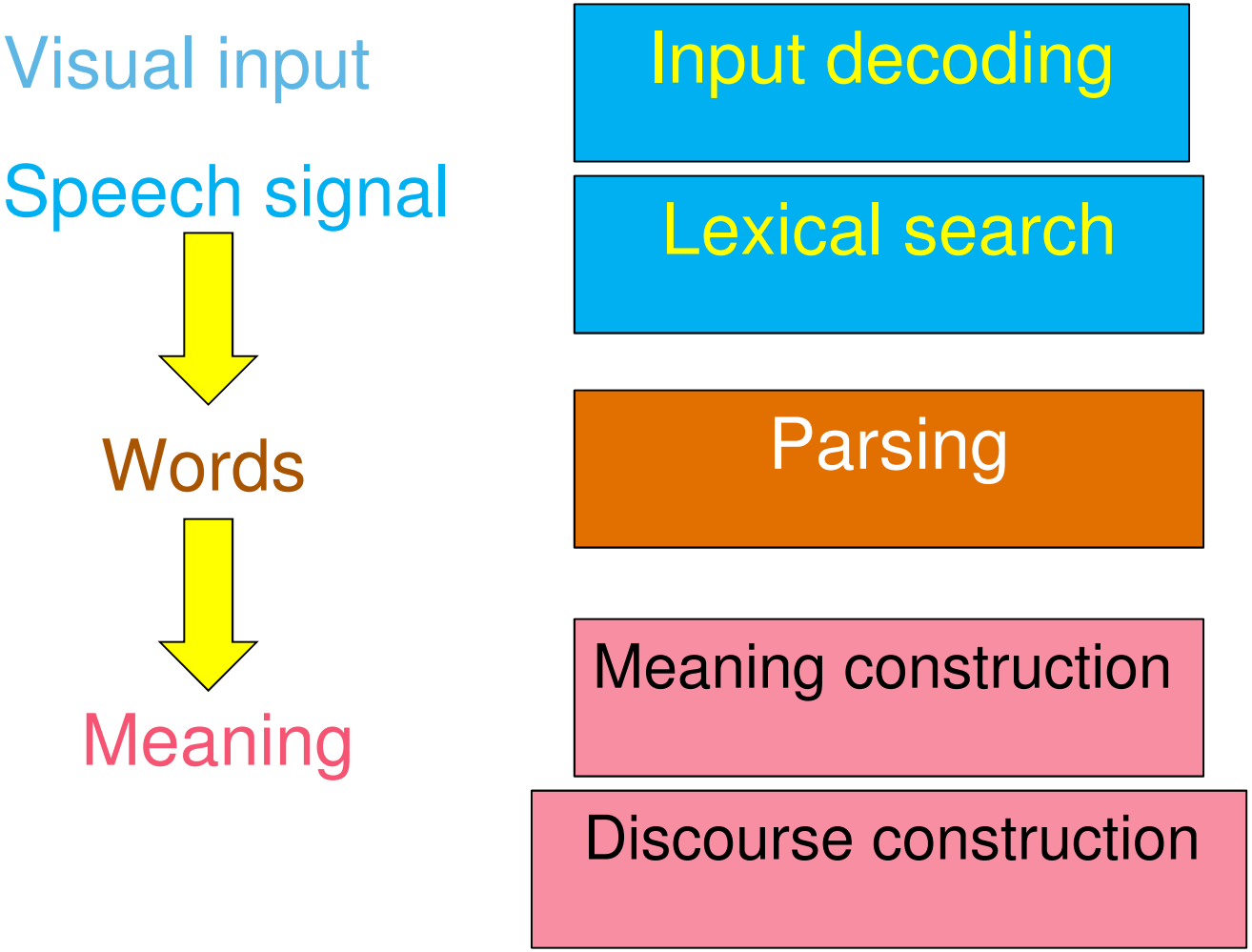
→ Studying candidate behaviour (verbal report)

University of Bedfordshire

Modelling the skill

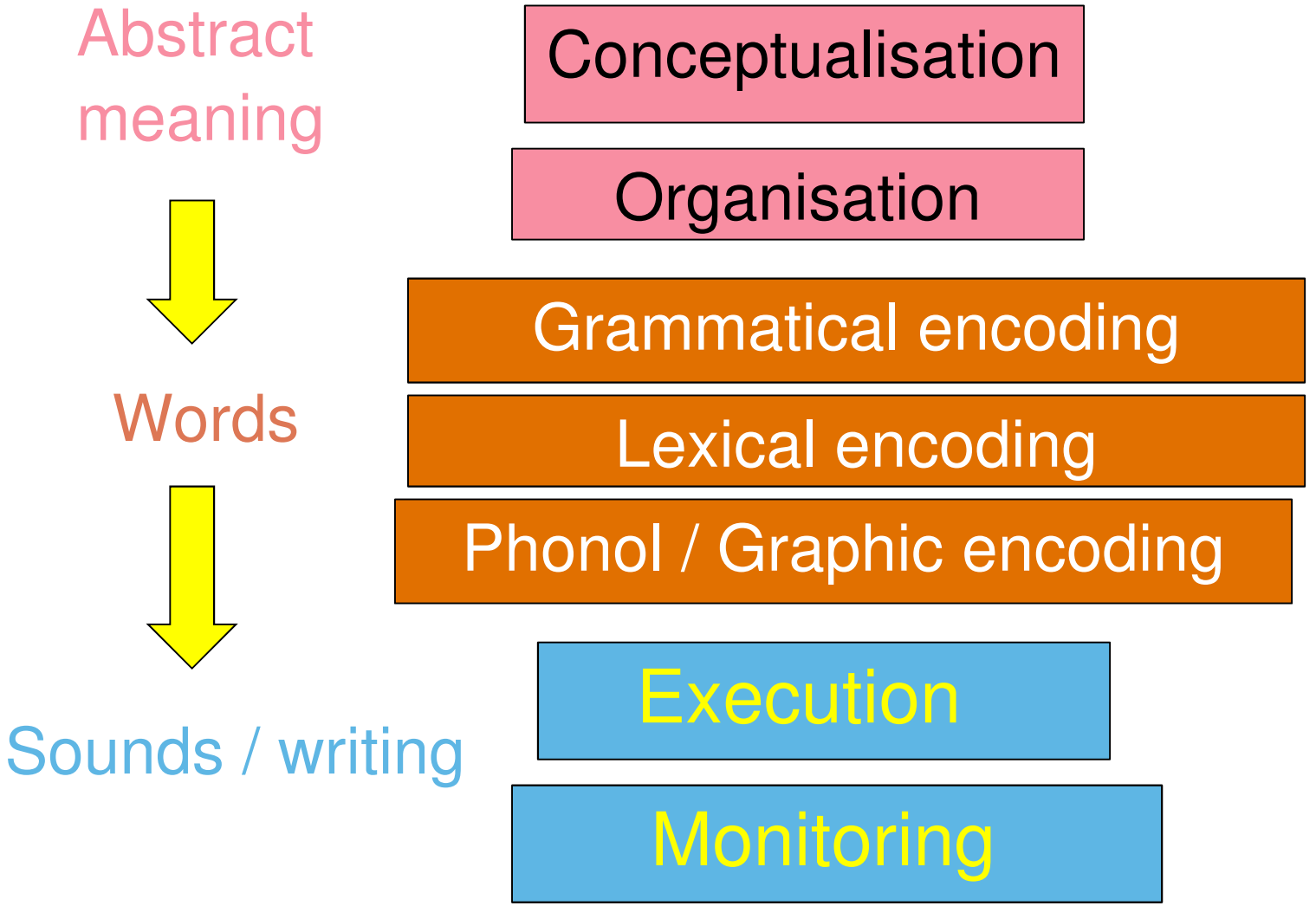
Phases of receptive skills

(Field 2008/2013)



Phases of productive skills (Levelt, 1989, 1999, Field 2011, Kellogg, 1996, Shaw & Weir, 2009)

University of Bedfordshire



Input constraints

University of Bedfordshire

- Writing and speaking require conceptual input (provided by test or provided by candidate?)
- The input in listening is highly variable at phoneme, word and sentence level - also between speakers and even within the speech of a single speaker. Word boundaries are not clearly marked. By contrast, reading has a standardised spelling and fonts, punctuation, and gaps between words
- Readers can check their understanding; listeners have to carry forward a discourse representation in their minds, which is often approximate.
- Should we expect a more general reporting of meaning in a listening test?

Performance constraints

University of Bedfordshire

- Listening and most speaking events are time-constrained, with the language user obliged to process language spontaneously. [But pre-planning may be factored in to a speaking test.]
- Writing is highly planned and recursive.
- Listening takes place at a pace determined by the speaker, while the pace of reading is determined by the reader (in response to subject matter and goals).
- Should we expect a more general reporting of meaning in a listening test?

University of Bedfordshire

A cognitive validation exercise

Three cognitive validity questions

University of Bedfordshire

- 1. To what extent are the cognitive processes elicited by a test comparable to those that would be employed in a real-world setting?
- 2. Is the range of processes elicited by a test comprehensive enough to be representative of behaviour in a real-world setting?
- 3. Are the cognitive demands imposed by a test sufficiently finely calibrated to reflect the level of the test?

University of Bedfordshire

1. Are the cognitive processes comparable?

Test format conventions

University of Bedfordshire

- Items are presented before listening. They
  - provide more information than would normally be available ahead of listening (and provide it in a different modality)
  - encourage the candidate to anticipate what will be heard (sometimes incorrectly)
- The need to read and internalise the items dictates that
  - items have to be presented in the same order as the passage.
  - Items have to be spaced out

MCQ sample

University of Bedfordshire

You hear an explorer talking about a journey he's making. How will he travel once he is across the river?

A. by motor vehicle

B. on horseback

C. on foot

(*FCE Handbook*, 2008: 60)

MCQ processing

(FCE Sample Test 1:1)

University of Bedfordshire

- The engine's full of water at the moment, it's very doubtful if any of the trucks can get across the river in this weather. The alternative is to carry all the stuff across using the old footbridge, which is perfectly possible …and then use horses rather than trucks for the rest of the trip all the way instead of just the last 10 or 15 kilometres as was our original intention. We can always pick up the vehicles again on the way back down…

MCQ processing

(FCE Sample Test 1:1)

University of Bedfordshire

- The engine's full of water at the moment, it's very doubtful if any of the trucks can get across the river in this weather. The alternative is to carry all the stuff across using the old footbridge, which is perfectly possible …and then use horses rather than trucks for the rest of the trip all the way instead of just the last 10 or 15 kilometres as was our original intention. We can always pick up the vehicles again on the way down…

MCQ processing

(FCE Sample Test 1:1))

- [① The engine's full of water at the moment], [② it's very doubtful if any of the trucks can get across the river in this weather]. [③ The alternative is to carry all the stuff across using the old footbridge], [④ which is perfectly possible] [⑤…and then [⑥ use horses rather than trucks for the rest of the trip] [⑦ all the way instead of just the last 10 or 15 kilometres] [⑧ as was the original intention]. [⑨ We can always pick up the vehicles again on the way down] [⑩…]

Processing of test formats

University of Bedfordshire

- The test formats used in listening tests are chosen because of their reliability and ease of marking. BUT
- They impose quite heavy cognitive demands upon the candidate who has to:
  - Internalise information from the items
  - Map from the items to the listening passage (which the items often paraphrase)
  - Decide how closely each new idea in the listening passage fits the current item
  - (MCQ) eliminate options that are negated in the recording.
- These operations are more demanding than normal listening

University of Bedfordshire

2. Is the range of processes elicited comprehensive enough?

Discourse construction

University of Bedfordshire

Choose	**Is it important? Is it relevant?**
Connect	**How is it linked to the last utterance?**
Compare	**Is it consistent with what was said so far?**
Construct	**What is the overall line of argument?**

Discourse construction overlooked

University of Bedfordshire

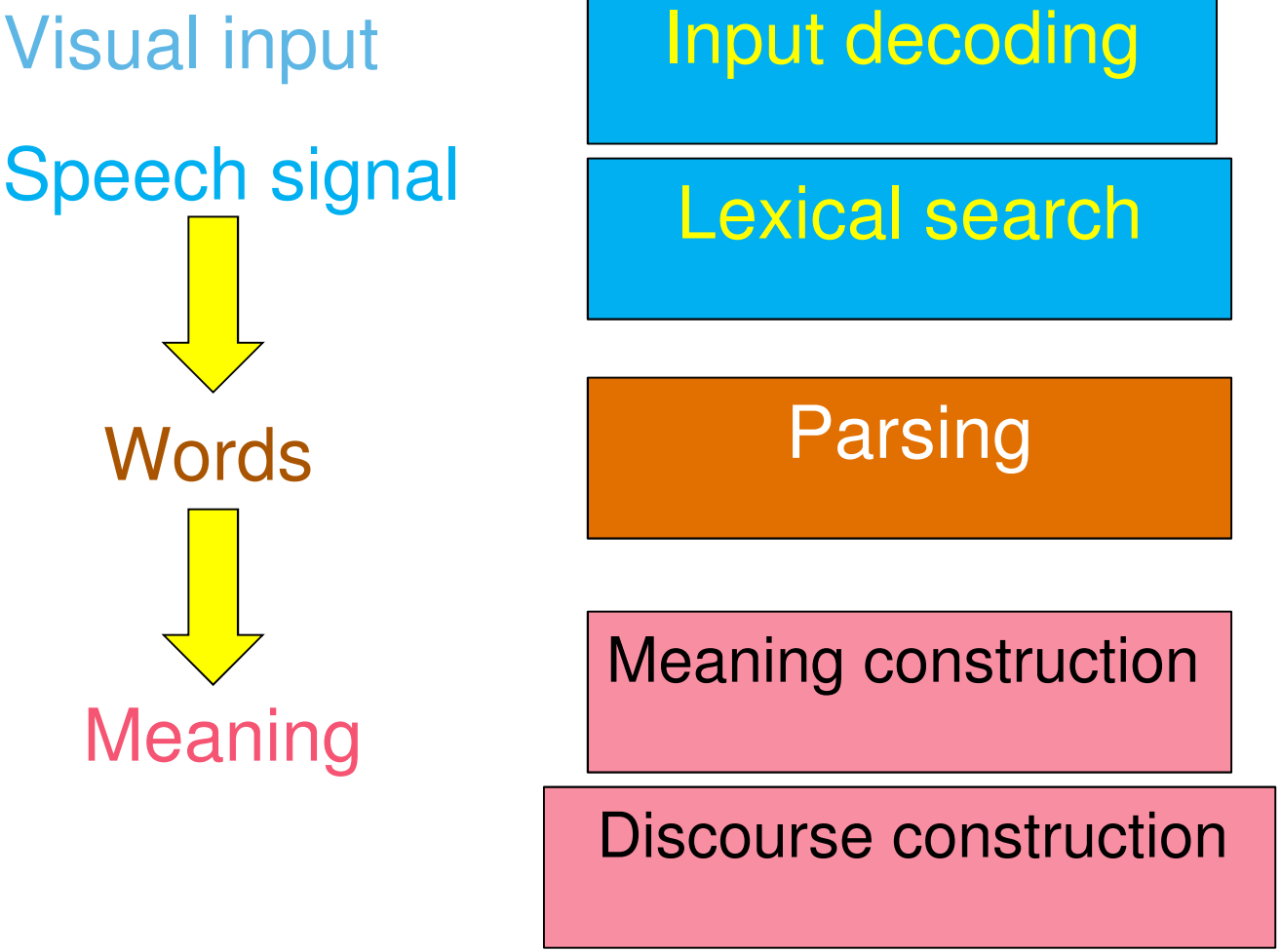
- Choose: the tester chooses which information points to focus on – sometimes choosing points that are not central to the recording
- Connect: Much testing focuses on single points, with no connection to those before and after
- Compare: Tests rarely ask learners to check information (for example, comparing two accounts of an accident)
- Construct. Tests rarely seek for evidence that learners can construct an outline based upon macro-and micro points / headings and subheadings

University of Bedfordshire

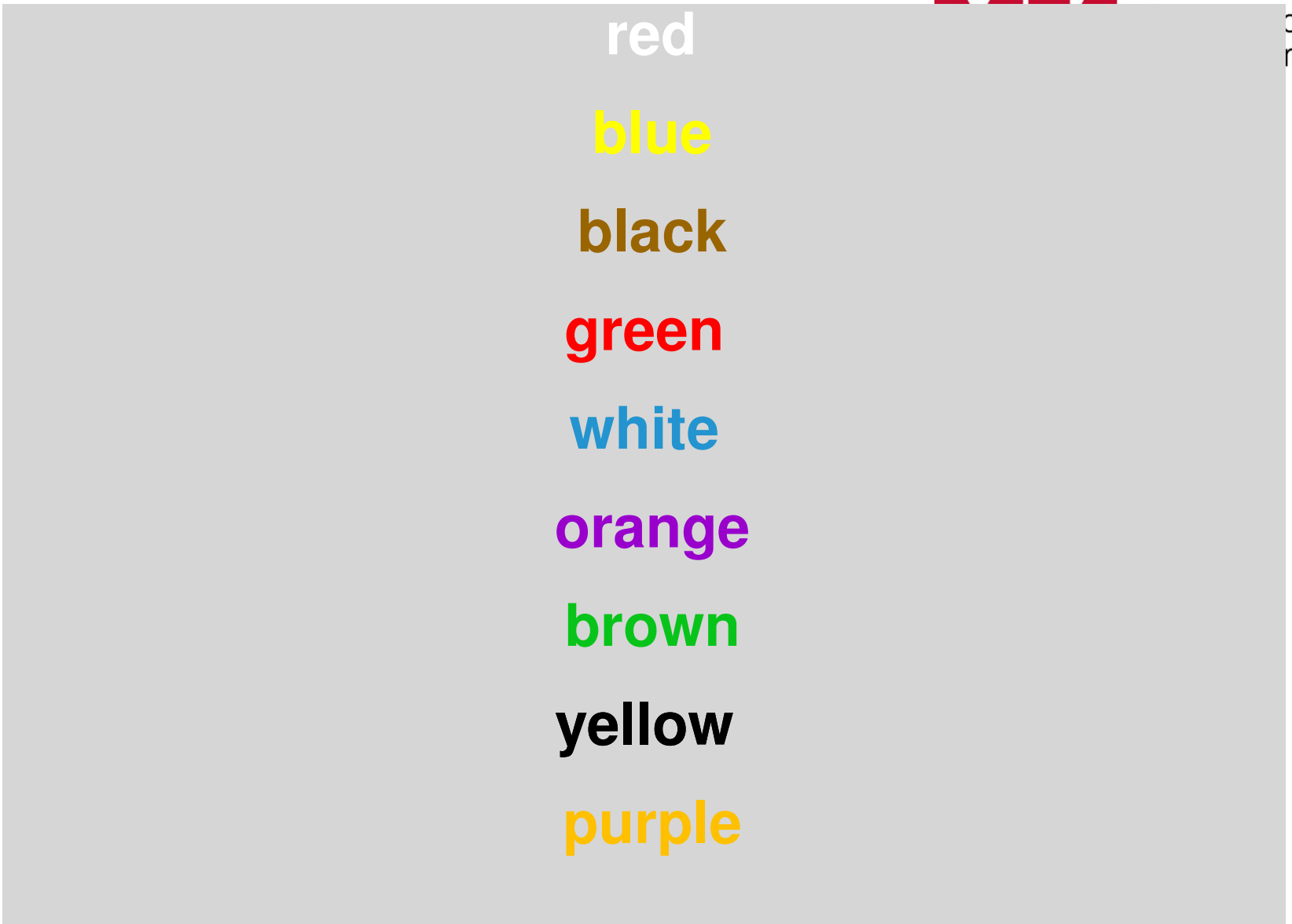
3. Are the cognitive demands made of test takers finely enough calibrated across levels?

Phases of receptive skills

(Field 2008/2013)



Automaticity: the Stroop test (Stroop, 1935)



Assumptions

University of Bedfordshire

- At lower levels, test takers cannot process the L2 with a sufficient degree of automaticity.
- In listening / reading, decoding, lexical search and parsing demand heavy resources of attention. This limits the extent to which test takers can be expected to process at higher levels (meaning construction / discourse construction)
- In speaking / writing, lexical, grammatical and phonological encoding demand heavy resources of attention.This limits the extent to which test takers can given attention to pragmatics, style, and register

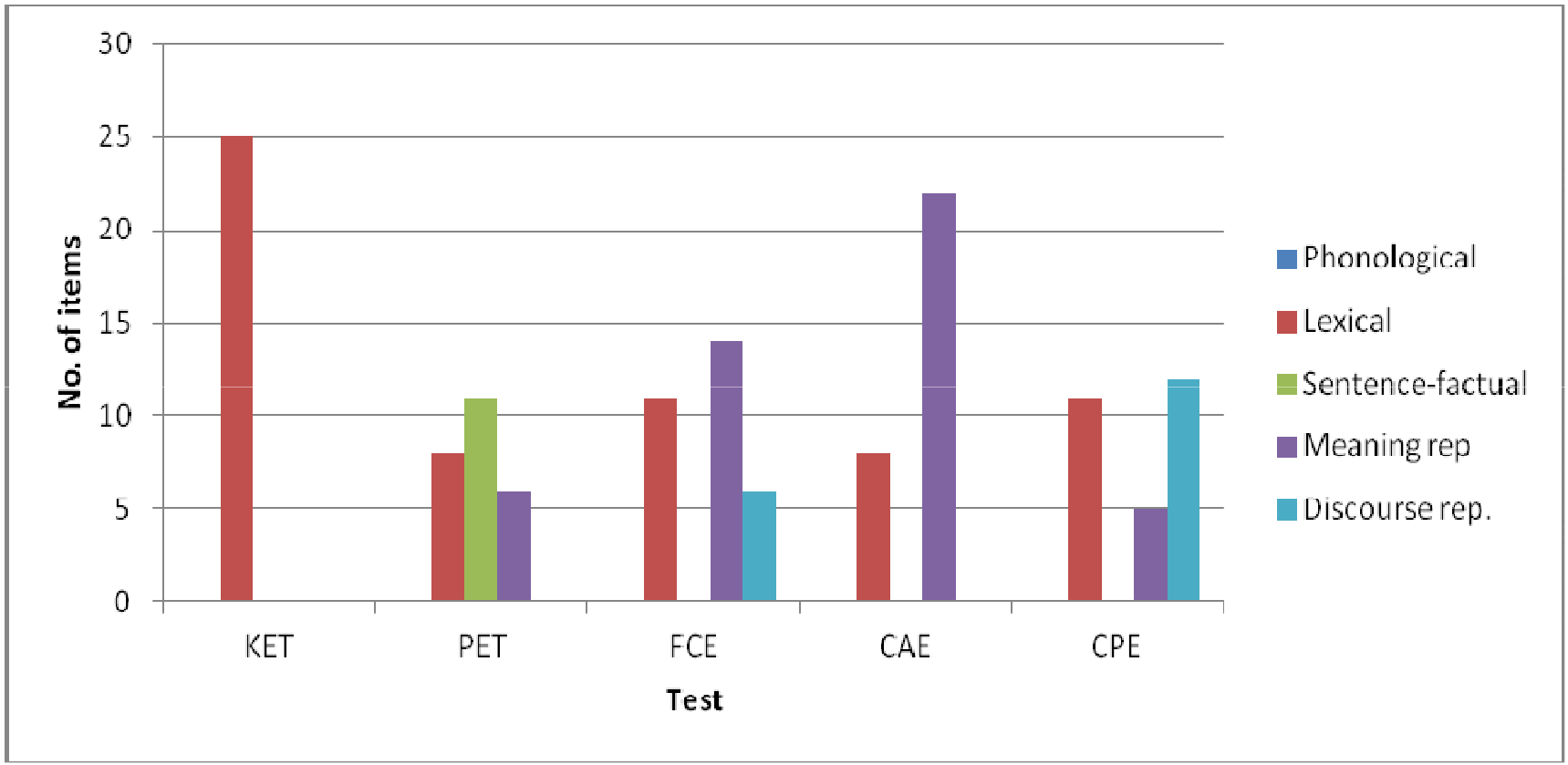
Calibration of processing

University of Bedfordshire

- Sample listening tests across five levels of the Cambridge ESOL suite
- Analysed in terms of the information focus: the unit of language that was targeted (e.g. word / lexical chunk, clause, inference not directly expressed, discourse relationship).
- This was taken to roughly indicate the highest level at which processing was demanded of the test taker (lexical search – parsing – meaning construction – discourse construction).
- Number of items for each unit was calculated.

Items tapping into various information units

University of Bedfordshire



Conclusions

University of Bedfordshire

- Assumptions confirmed. At lower proficiency level, items tap into smaller and less cognitively complex information units – entailing lexical search and parsing (= factual information)
- No coverage of discrimination at phonological level – a discredited approach to listening after early attempts to test the skill through phoneme discrimination.
- Lexical search features across levels – mainly the effect of using the gap filling format
- Low coverage of discourse representation
- Simple factual information not represented at FCE level.

One reservation…

University of Bedfordshire

- While test takers at lower proficiency levels have limited resources of attention to allocate to higher level processing, they should still be capable of using compensatory strategies in order to infer the main point of a listening passage, despite local difficulties in decoding words and phrases.
- Provision should be made at lower levels for items that require test takers to report the main point / gist.

University of Bedfordshire

Examining candidate behaviour

Approach 1.
Three types of behaviour (Field, 2011)

University of Bedfordshire

- 1 Part of the normal process: behaviour which might be adopted by an L1 listener.
- 2. Strategic behaviour to prepare for a task, to maximise the amount retained or to compensate for problems of understanding.
- 3 Task-specific behaviour representing the user's response to features of the task.
    - a. processes related to the task but not part of the corresponding real-life activity
    - b. strategies where the learner attempts to exploit loopholes in the format of the task

Test-wise strategies (IELTS study)

University of Bedfordshire

Test- wise strategies employing visual cues

- **Q match**. Listened for words in the spoken text that formed a one- to-one match with words in the written
- **Q loc.** Used a word or words from the written text to locate information in the spoken text
- **Q para**. Sought a paraphrase in the spoken text of a proposition expressed in the written one
- **Q seq:** Chose an answer according to its position in a list or in a sequence of propositions in the written test

Test wise strategies with visual cues (% of all responses)

University of Bedfordshire

	Q match	Q loc	Q para	Q seq
Test A (N = 13)	0	26 (18.98%)	2 (1.46%)	15 (10.95%)
Test B (N = 13)	30 (22.56%)	25 (18.80%)	2 (1.50%)	3 (2.56%)

Test A: gap filling

Test B: gap filling and MCQ

Approach 2.
Comparing candidate behaviour

University of Bedfordshire

- A. Across conditions. (Field 2011)
- Test taker behaviour / results in test conditions compared with behaviour / results of same individuals in conditions closer to real world ones (e.g listening to a lecture and note-taking).
- Test-takers asked to compare relative difficulty
- B. Across populations
- L2 test taker behaviour in test conditions compared with L1 test taker behaviour under same conditions

Findings (IELTS project)

- No correlation between scores in test and non-test conditions
- 8 of 28 participants categorically asserted that they found lecture-style listening and note-taking easier than operating under test conditions.

Conclusion

University of Bedfordshire

Cognitive Validity research and validation:

- 1. compares what we know from empirical findings about the processes that contribute to a target construct
- .against the input to the test taker
- against the formats used
- against item content
- 2. compares the behaviour of the test taker under test conditions and under those that more closely replicate real world language use.

John.Field@beds.ac.uk